Loss function landscape. Part 2

Theories of Deep Learning

Eugene Golikov

MIPT, spring 2019

Neural Networks and Deep Learning Lab., MIPT

Linear nets

Deep linear net:

$$\mathcal{L}_{net}(W) = \|Y - W_H W_{H-1} \dots W_1 X\|_F^2 \to \min_W .$$

All local minima are global.

Linear nets

Deep linear net:

$$\mathcal{L}_{net}(W) = \|Y - W_H W_{H-1} \dots W_1 X\|_F^2 \to \min_W .$$

All local minima are global.

However, there exist "bad saddles" with no negative values of the Hessian (Kawaguchi, 2016), e.g.:

$\nabla \mathcal{L}_{net}(\mathbf{0}) = 0$ and $\nabla^2 \mathcal{L}_{net}(\mathbf{0}) = 0$ for $H \geq 3$.

Linear nets

Let $d_0 = d_1 = \ldots = d_H$;
Let $y = Rx + \xi$, where $\xi \sim \mathcal{N}(0, I)$.

Let's reparameterize our linear net as a ResNet:

$$\mathcal{L}_{resnet}(W) = \mathbb{E}\|y - (I + W_H)(I + W_{H-1})\ldots(I + W_1)x\|_2^2 \to \min_W.$$

[1] https://arxiv.org/abs/1611.04231

Linear nets

Let $d_0 = d_1 = \ldots = d_H$;
Let $y = Rx + \xi$, where $\xi \sim \mathcal{N}(0, I)$.

Let's reparameterize our linear net as a ResNet:

$$\mathcal{L}_{resnet}(W) = \mathbb{E}\|y - (I + W_H)(I + W_{H-1})\ldots(I + W_1)x\|_2^2 \to \min_W.$$

Theorem 1 (Hardt & Ma, 2016[1]):
Any critical point of $\mathcal{L}_{resnet}(W)$ for which $\max_{k=1,\ldots,H} \|W_k\| < 1$ is a global minimum.

[1] https://arxiv.org/abs/1611.04231

Linear nets

Let $d_0 = d_1 = \ldots = d_H$;
Let $y = Rx + \xi$, where $\xi \sim \mathcal{N}(0, I)$.

Let's reparameterize our linear net as a ResNet:

$$\mathcal{L}_{resnet}(W) = \mathbb{E}\|y - (I + W_H)(I + W_{H-1})\ldots(I + W_1)x\|_2^2 \to \min_W.$$

Theorem 1 (Hardt & Ma, 2016[1]):

Any critical point of $\mathcal{L}_{resnet}(W)$ for which $\max_{k=1,\ldots,H} \|W_k\| < 1$ is a global minimum.

Theorem 2 (Hardt & Ma, 2016):

There is a sequence $\{W^{(H)}\}_{H=1}^{\infty}$ of global minima with $\lim_{H\to\infty} \max_{k=1,\ldots,H} \|W_k^{(H)}\| = 0$.

[1] https://arxiv.org/abs/1611.04231

Shallow non-linear nets

The simplest non-linear net:

$$\hat{y}(x, w) = \sigma(w^T x);$$

$$\mathcal{L}(w) = \mathbb{E}_{x,y\sim\mathcal{D}}(y - \sigma(w^T x))^2.$$

Some known results:

[2] https://papers.nips.cc/paper/1028-exponentially-many-local-minima-for-single-neurons

[3] https://arxiv.org/abs/1703.00560

Shallow non-linear nets

The simplest non-linear net:

$$\hat{y}(x, w) = \sigma(w^T x);$$

$$\mathcal{L}(w) = \mathbb{E}_{x,y\sim\mathcal{D}}(y - \sigma(w^T x))^2.$$

Some known results:

- Auer et al. (1995)[2]: if $\sigma(z) = (1 + \exp(-z))^{-1}$, then there exists a finite dataset, for which there are multiple local minima of $\mathcal{L}$;

[2] https://papers.nips.cc/paper/1028-exponentially-many-local-minima-for-single-neurons

[3] https://arxiv.org/abs/1703.00560

Shallow non-linear nets

The simplest non-linear net:

$$\hat{y}(x, w) = \sigma(w^T x);$$

$$\mathcal{L}(w) = \mathbb{E}_{x,y\sim\mathcal{D}}(y - \sigma(w^T x))^2.$$

Some known results:

- Auer et al. (1995)[2]: if $\sigma(z) = (1 + \exp(-z))^{-1}$, then there exists a finite dataset, for which there are multiple local minima of $\mathcal{L}$;
- Tian (2017)[3]: if $\sigma(z) = [z]_+$, $x \sim \mathcal{N}(0, I)$, $y(x) = \sigma(w_*^T x)$, then $w = w_*$ is a unique minimum of $\mathcal{L}(w)$.

[2] https://papers.nips.cc/paper/1028-exponentially-many-local-minima-for-single-neurons

[3] https://arxiv.org/abs/1703.00560

Shallow non-linear nets

A bit more complex non-linear net:

$$\mathcal{L}(W) = \mathbb{E}_{x,y\sim\mathcal{D}} \left(y - \sum_{k=1}^{K} \sigma(w_k^T x) \right)^2 .$$

Some known results:

Consider $\sigma(z) = [z]_+$, $x \sim \mathcal{N}(0, I)$, $y(x) = \sum_{k=1}^{K} \sigma(w_{*,k}^T x)$;

[4] https://openreview.net/forum?id=B14uJzW0b

[5] https://arxiv.org/abs/1712.08968

Shallow non-linear nets

A bit more complex non-linear net:

$$\mathcal{L}(W) = \mathbb{E}_{x,y\sim\mathcal{D}} \left(y - \sum_{k=1}^{K} \sigma(w_k^T x) \right)^2 .$$

Some known results:

Consider $\sigma(z) = [z]_+$, $x \sim \mathcal{N}(0, I)$, $y(x) = \sum_{k=1}^{K} \sigma(w_{*,k}^T x)$;

- Tian (2017): for $K = 1$ $W = W_*$ is a unique minimum of $\mathcal{L}(W)$.

[4] https://openreview.net/forum?id=B14uJzW0b
[5] https://arxiv.org/abs/1712.08968

Shallow non-linear nets

A bit more complex non-linear net:

$$\mathcal{L}(W) = \mathbb{E}_{x,y\sim\mathcal{D}} \left(y - \sum_{k=1}^{K} \sigma(w_k^T x) \right)^2 .$$

Some known results:

Consider $\sigma(z) = [z]_+$, $x \sim \mathcal{N}(0, I)$, $y(x) = \sum_{k=1}^{K} \sigma(w_{*,k}^T x)$;

- Tian (2017): for $K = 1$ $W = W_*$ is a unique minimum of $\mathcal{L}(W)$.
- Wu et al. (2018)[4]: for $K = 2$, if $w_{*,1} \perp w_{*,2}$, $\|w_{*,1}\| = \|w_{*,2}\| = 1$, $W = W_*$ is a unique minimum of $\mathcal{L}(W)$ of norm 1;

[4] https://openreview.net/forum?id=B14uJzW0b
[5] https://arxiv.org/abs/1712.08968

Shallow non-linear nets

A bit more complex non-linear net:

$$\mathcal{L}(W) = \mathbb{E}_{x,y\sim\mathcal{D}} \left(y - \sum_{k=1}^{K} \sigma(w_k^T x) \right)^2 .$$

Some known results:

Consider $\sigma(z) = [z]_+$, $x \sim \mathcal{N}(0, I)$, $y(x) = \sum_{k=1}^{K} \sigma(w_{*,k}^T x)$;

- Tian (2017): for $K = 1$ $W = W_*$ is a unique minimum of $\mathcal{L}(W)$.
- Wu et al. (2018)[4]: for $K = 2$, if $w_{*,1} \perp w_{*,2}$, $\|w_{*,1}\| = \|w_{*,2}\| = 1$, $W = W_*$ is a unique minimum of $\mathcal{L}(W)$ of norm 1;
- Safran & Shamir (2017)[5]: for $6 \leq K \leq 20$ there are multiple non-global minima of $\mathcal{L}$.

[4] https://openreview.net/forum?id=B14uJzW0b
[5] https://arxiv.org/abs/1712.08968

Shallow non-linear nets

A non-linear net with one hidden layer:

$$\mathcal{L}(W) = \|Y - W_2\sigma(W_1X)\|_F^2,$$

where $X \in \mathbb{R}^{d_0 \times m}$, $W_1 \in \mathbb{R}^{d_1 \times d_0}$, $W_2 \in \mathbb{R}^{d_2 \times d_1}$ and $Y \in \mathbb{R}^{d_2 \times m}$.

Theorem Yu & Chen (1995)[6]:

[6] https://ieeexplore.ieee.org/document/410380

Shallow non-linear nets

A non-linear net with one hidden layer:

$$\mathcal{L}(W) = \|Y - W_2\sigma(W_1X)\|_F^2,$$

where $X \in \mathbb{R}^{d_0 \times m}$, $W_1 \in \mathbb{R}^{d_1 \times d_0}$, $W_2 \in \mathbb{R}^{d_2 \times d_1}$ and $Y \in \mathbb{R}^{d_2 \times m}$.

Theorem Yu & Chen (1995)[6]:

Suppose

1. $\sigma(z) = (1 + \exp(-z))^{-1}$,
2. all columns of X are distinct,
3. $d_1 = m$.

[6] https://ieeexplore.ieee.org/document/410380

Shallow non-linear nets

A non-linear net with one hidden layer:

$$\mathcal{L}(W) = \|Y - W_2\sigma(W_1X)\|_F^2,$$

where $X \in \mathbb{R}^{d_0 \times m}$, $W_1 \in \mathbb{R}^{d_1 \times d_0}$, $W_2 \in \mathbb{R}^{d_2 \times d_1}$ and $Y \in \mathbb{R}^{d_2 \times m}$.

Theorem Yu & Chen (1995)[6]:

Suppose

1. $\sigma(z) = (1 + \exp(-z))^{-1}$,
2. all columns of X are distinct,
3. $d_1 = m$.

Then all local minima of $\mathcal{L}$ are global.

[6] https://ieeexplore.ieee.org/document/410380

Deep non-linear nets

A non-linear net with multiple hidden layers:

$$\mathcal{L}(W_{1:H}) = \|Y - W_H\sigma(W_{H-1}\dots\sigma(W_1x)\dots)\|_F^2,$$

where $X \in \mathbb{R}^{d_0\times m}$, $\forall k\ W_k \in \mathbb{R}^{d_k\times d_{k-1}}$ and $Y \in \mathbb{R}^{d_H\times m}$.

Theorem (Nguyen & Hein (2017)[7]):

Suppose $\sigma(z) = (1+\exp(-z))^{-1}$ and all columns of X are distinct.

[7] https://arxiv.org/abs/1704.08045

Deep non-linear nets

A non-linear net with multiple hidden layers:

$$\mathcal{L}(W_{1:H}) = \|Y - W_H\sigma(W_{H-1}\dots\sigma(W_1x)\dots)\|_F^2,$$

where $X \in \mathbb{R}^{d_0\times m}$, $\forall k\ W_k \in \mathbb{R}^{d_k\times d_{k-1}}$ and $Y \in \mathbb{R}^{d_H\times m}$.

Theorem (Nguyen & Hein (2017)[7]):

Suppose $\sigma(z) = (1+\exp(-z))^{-1}$ and all columns of X are distinct.
Let $W^*_{1:H}$ be a local minimum; if following conditions hold:

1. $\exists k:\ d_k \geq m,\ \forall l > k+1 \quad \operatorname{rk} W_l = d_l$,
2. hessian of $\mathcal{L}$ wrt $W_{k+1:H}$ is non-degenerate at $W^*_{1:H}$,

[7] https://arxiv.org/abs/1704.08045

Deep non-linear nets

A non-linear net with multiple hidden layers:

$$\mathcal{L}(W_{1:H}) = \|Y - W_H\sigma(W_{H-1}\dots\sigma(W_1x)\dots)\|_F^2,$$

where $X \in \mathbb{R}^{d_0\times m}$, $\forall k\ W_k \in \mathbb{R}^{d_k\times d_{k-1}}$ and $Y \in \mathbb{R}^{d_H\times m}$.

Theorem (Nguyen & Hein (2017)[7]):

Suppose $\sigma(z) = (1+\exp(-z))^{-1}$ and all columns of X are distinct. Let $W^*_{1:H}$ be a local minimum; if following conditions hold:

1. $\exists k:\ d_k \geq m,\ \forall l > k+1 \quad \operatorname{rk} W_l = d_l$,
2. hessian of $\mathcal{L}$ wrt $W_{k+1:H}$ is non-degenerate at $W^*_{1:H}$,

then $W^*_{1:H}$ is a global minimum.

[7] https://arxiv.org/abs/1704.08045